Advanced High School Statistics

Preliminary Edition

Chapter 2

Summarizing Data

2.3.1 Contingency tables and bar plots

Table 2.26 summarizes two variables: `spam` and `number`. Recall that `number` is a categorical variable that describes whether an email contains no numbers, only small numbers (values under 1 million), or at least one big number (a value of 1 million or more). A table that summarizes data for two categorical variables in this way is called a **contingency table**. Each value in the table represents the number of times a particular combination of variable outcomes occurred. For example, the value 149 corresponds to the number of emails in the data set that are spam *and* had no number listed in the email. Row and column totals are also included. The **row totals** provide the total counts across each row (e.g. $149 + 168 + 50 = 367$), and **column totals** are total counts down each column.

Table 2.27 shows a frequency table for the `number` variable. If we replaced the counts with percentages or proportions, the table is a **relative frequency table**.

		`number`			
		none	small	big	Total
`spam`	spam	149	168	50	367
	not spam	400	2659	495	3554
	Total	549	2827	545	3921

Table 2.26: A contingency table for `spam` and `number`.

none	small	big	Total
549	2827	545	3921

Table 2.27: A frequency table for the `number` variable.

Because the numbers in these tables are counts, not to data points, they cannot be graphed using the methods we applied to numerical data. Instead, another set of graphing methods are needed that are suitable for categorical data.

A bar plot is a common way to display a single categorical variable. The left panel of Figure 2.28 shows a **bar plot** for the `number` variable. In the right panel, the counts are converted into proportions (e.g. $549/3921 = 0.140$ for `none`), showing the proportion of observations that are in each level (i.e. in each category).

2.3.2 Row and column proportions

Table 2.29 shows the row proportions for Table 2.26. The **row proportions** are computed as the counts divided by their row totals. The value 149 at the intersection of `spam` and `none` is replaced by $149/367 = 0.406$, i.e. 149 divided by its row total, 367. So what does 0.406 represent? It corresponds to the proportion of spam emails in the sample that do not have any numbers.

A contingency table of the column proportions is computed in a similar way, where each **column proportion** is computed as the count divided by the corresponding column total. Table 2.30 shows such a table, and here the value 0.271 indicates that 27.1% of emails with no numbers were spam. This rate of spam is much higher compared to emails with only small numbers (5.9%) or big numbers (9.2%). Because these spam rates vary between the three levels of `number` (`none`, `small`, `big`), this provides evidence that the `spam` and `number` variables are associated.

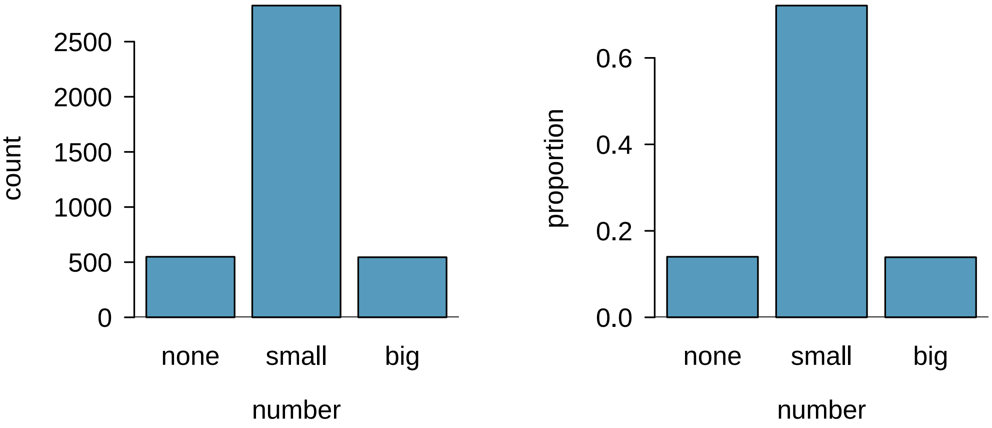


Figure 2.28: Two bar plots of `number`. The left panel shows the counts, and the right panel shows the proportions in each group.

	none	small	big	Total
spam	$149/367 = 0.406$	$168/367 = 0.458$	$50/367 = 0.136$	1.000
not spam	$400/3554 = 0.113$	$2657/3554 = 0.748$	$495/3554 = 0.139$	1.000
Total	$549/3921 = 0.140$	$2827/3921 = 0.721$	$545/3921 = 0.139$	1.000

Table 2.29: A contingency table with row proportions for the `spam` and `number` variables.

We could also have checked for an association between `spam` and `number` in Table 2.29 using row proportions. When comparing these row proportions, we would look down columns to see if the fraction of emails with no numbers, small numbers, and big numbers varied from `spam` to `not spam`.

Guided Practice 2.52 What does 0.458 represent in Table 2.29? What does 0.059 represent in Table 2.30?[36]

Guided Practice 2.53 What does 0.139 at the intersection of `not spam` and `big` represent in Table 2.29? What does 0.908 represent in the Table 2.30?[37]

	none	small	big	Total
spam	149/549 = 0.271	168/2827 = 0.059	50/545 = 0.092	367/3921 = 0.094
not spam	400/549 = 0.729	2659/2827 = 0.941	495/545 = 0.908	3684/3921 = 0.906
Total	1.000	1.000	1.000	1.000

Table 2.30: A contingency table with column proportions for the `spam` and `number` variables.

Example 2.54 Data scientists use statistics to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is whether or not an email has any HTML content. A contingency table for the `spam` and `format` variables from the `email` data set are shown in Table 2.31. Recall that an HTML email is an email with the capacity for special formatting, e.g. bold text. In Table 2.31, which would be more helpful to someone hoping to classify email as spam or regular email: row or column proportions?

Such a person would be interested in how the proportion of spam changes within each email format. This corresponds to column proportions: the proportion of spam in plain text emails and the proportion of spam in HTML emails.

If we generate the column proportions, we can see that a higher fraction of plain text emails are spam (209/1195 = 17.5%) than compared to HTML emails (158/2726 = 5.8%). This information on its own is insufficient to classify an email as spam or not spam, as over 80% of plain text emails are not spam. Yet, when we carefully combine this information with many other characteristics, such as `number` and other variables, we stand a reasonable chance of being able to classify some email as spam or not spam.

	text	HTML	Total
spam	209	158	367
not spam	986	2568	3554
Total	1195	2726	3921

Table 2.31: A contingency table for `spam` and `format`.

[36] 0.458 represents the proportion of spam emails that had a small number. 0.058 represents the fraction of emails with small numbers that are spam.

[37] 0.139 represents the fraction of non-spam email that had a big number. 0.908 represents the fraction of emails with big numbers that are non-spam emails.

Example 2.54 points out that row and column proportions are not equivalent. Before settling on one form for a table, it is important to consider each to ensure that the most useful table is constructed.

⊙ **Guided Practice 2.55** Look back to Tables 2.29 and 2.30. Which would be more useful to someone hoping to identify spam emails using the `number` variable?[38]

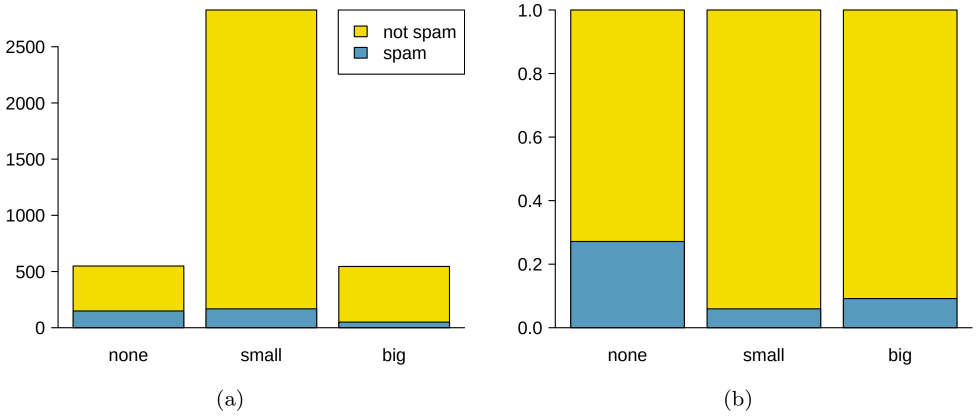


Figure 2.32: (a) Segmented bar plot for numbers found in emails, where the counts have been further broken down by `spam`. (b) Standardized version of Figure (a).

[38] The column proportions in Table 2.30 will probably be most useful, which makes it easier to see that emails with small numbers are spam about 5.9% of the time (relatively rare). We would also see that about 27.1% of emails with no numbers are spam, and 9.2% of emails with big numbers are spam.

2.3.3 Segmented bar plots

Contingency tables using row or column proportions are especially useful for examining how two categorical variables are related. Segmented bar plots provide a way to visualize the information in these tables.

A **segmented bar plot** is a graphical display of contingency table information. For example, a segmented bar plot representing Table 2.30 is shown in Figure 2.32(a), where we have first created a bar plot using the `number` variable and then divided each group by the levels of `spam`. The column proportions of Table 2.30 have been translated into a standardized segmented bar plot in Figure 2.32(b), which is a helpful visualization of the fraction of spam emails in each level of `number`.

● **Example 2.56** Examine both of the segmented bar plots. Which is more useful?

Figure 2.32(a) contains more information, but Figure 2.32(b) presents the information more clearly. This second plot makes it clear that emails with no number have a relatively high rate of spam email – about 27%! On the other hand, less than 10% of email with small or big numbers are spam.

Since the proportion of spam changes across the groups in Figure 2.32(b), we can conclude the variables are dependent, which is something we were also able to discern using table proportions. Because both the `none` and `big` groups have relatively few observations compared to the `small` group, the association is more difficult to see in Figure 2.32(a).

In some other cases, a segmented bar plot that is not standardized will be more useful in communicating important information. Before settling on a particular segmented bar plot, create standardized and non-standardized forms and decide which is more effective at communicating features of the data.

Chapter 3

Probability

3.1.5 Independence

Just as variables and observations can be independent, random processes can be independent, too. Two processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other. For instance, flipping a coin and rolling a die are two independent processes – knowing the coin was heads does not help determine the outcome of a die roll. On the other hand, stock prices usually move up or down together, so they are not independent.

Example 3.5 provides a basic example of two independent processes: rolling two dice. We want to determine the probability that both will be `1`. Suppose one of the dice is red and the other white. If the outcome of the red die is a `1`, it provides no information about the outcome of the white die. We first encountered this same question in Example 3.5 (page 84), where we calculated the probability using the following reasoning: $1/6^{th}$ of the time the red die is a `1`, and $1/6^{th}$ of *those* times the white die will also be `1`. This is illustrated in Figure 3.6. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer: $(1/6) \times (1/6) = 1/36$. This can be generalized to many independent processes.

[14] Brief solutions: (a) $A^c = \{3, 4, 5, 6\}$ and $B^c = \{1, 2, 3, 5\}$. (b) Noting that each outcome is disjoint, add the individual outcome probabilities to get $P(A^c) = 2/3$ and $P(B^c) = 2/3$. (c) A and A^c are disjoint, and the same is true of B and B^c. Therefore, $P(A) + P(A^c) = 1$ and $P(B) + P(B^c) = 1$.

[15] (a) The complement of getting at least one 6 in ten rolls of a die is getting zero 6's in the 10 rolls. (b) The complement of getting at most three 6's in 10 rolls is getting four, five, ..., nine, or ten 6's in 10 rolls.

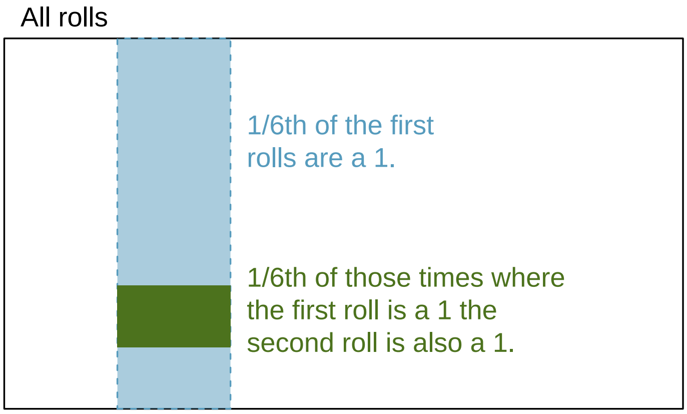

All rolls
1/6th of the first rolls are a 1.
1/6th of those times where the first roll is a 1 the second roll is also a 1.

Figure 3.6: $1/6^{th}$ of the time, the first roll is a `1`. Then $1/6^{th}$ of *those* times, the second roll will also be a `1`.

● **Example 3.25** What if there was also a blue die independent of the other two? What is the probability of rolling the three dice and getting all `1`s?

The same logic applies from Example 3.5. If $1/36^{th}$ of the time the white and red dice are both `1`, then $1/6^{th}$ of *those* times the blue die will also be `1`, so multiply:

$$\begin{aligned} P(white = \texttt{1} \text{ and } red = \texttt{1} \text{ and } blue = \texttt{1}) &= P(white = \texttt{1}) \times P(red = \texttt{1}) \times P(blue = \texttt{1}) \\ &= (1/6) \times (1/6) \times (1/6) = 1/216 \end{aligned}$$

Examples 3.5 and 3.25 illustrate what is called the Multiplication Rule for independent processes.

Multiplication Rule for independent processes

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B) \tag{3.26}$$

Similarly, if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

⊙ **Guided Practice 3.27** About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume these two people are independent. (a) What is the probability that both are left-handed? (b) What is the probability that both are right-handed?[16]

⊙ **Guided Practice 3.28** Suppose 5 people are selected at random.[17]

(a) What is the probability that all are right-handed?

(b) What is the probability that all are left-handed?

(c) What is the probability that not all of the people are right-handed?

Suppose the variables `handedness` and `gender` are independent, i.e. knowing someone's `gender` provides no useful information about their `handedness` and vice-versa. Then we can compute whether a randomly selected person is right-handed and female[18] using the Multiplication Rule:

$$\begin{aligned} P(\text{right-handed and female}) &= P(\text{right-handed}) \times P(\text{female}) \\ &= 0.91 \times 0.50 = 0.455 \end{aligned}$$

⊙ **Guided Practice 3.29** Three people are selected at random.[19]

(a) What is the probability that the first person is male and right-handed?

(b) What is the probability that the first two people are male and right-handed?.

(c) What is the probability that the third person is female and left-handed?

(d) What is the probability that the first two people are male and right-handed and the third person is female and left-handed?

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy Equation (3.26).

[16] (a) The probability the first person is left-handed is 0.09, which is the same for the second person. We apply the Multiplication Rule for independent processes to determine the probability that both will be left-handed: $0.09 \times 0.09 = 0.0081$.

(b) It is reasonable to assume the proportion of people who are ambidextrous (both right and left handed) is nearly 0, which results in $P(\text{right-handed}) = 1 - 0.09 = 0.91$. Using the same reasoning as in part (a), the probability that both will be right-handed is $0.91 \times 0.91 = 0.8281$.

[17] (a) The abbreviations `RH` and `LH` are used for right-handed and left-handed, respectively. Since each are independent, we apply the Multiplication Rule for independent processes:

$$\begin{aligned} P(\text{all five are } \texttt{RH}) &= P(\text{first} = \texttt{RH}, \text{ second} = \texttt{RH}, \text{ ..., fifth} = \texttt{RH}) \\ &= P(\text{first} = \texttt{RH}) \times P(\text{second} = \texttt{RH}) \times \cdots \times P(\text{fifth} = \texttt{RH}) \\ &= 0.91 \times 0.91 \times 0.91 \times 0.91 \times 0.91 = 0.624 \end{aligned}$$

(b) Using the same reasoning as in (a), $0.09 \times 0.09 \times 0.09 \times 0.09 \times 0.09 = 0.0000059$

(c) Use the complement, $P(\text{all five are } \texttt{RH})$, to answer this question:

$$P(\text{not all } \texttt{RH}) = 1 - P(\text{all } \texttt{RH}) = 1 - 0.624 = 0.376$$

[18] The actual proportion of the U.S. population that is `female` is about 50%, and so we use 0.5 for the probability of sampling a woman. However, this probability does differ in other countries.

[19] Brief answers are provided. (a) This can be written in probability notation as $P(\text{a randomly selected person is male and right-handed}) = 0.455$. (b) 0.207. (c) 0.045. (d) 0.0093.

● **Example 3.30** If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is 1/4 and the probability that it is an ace is 1/13. The probability the card is the ace of hearts is 1/52. We check whether Equation 3.26 is satisfied:

$$P(\heartsuit) \times P(\text{ace}) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52} = P(\heartsuit \text{ and ace})$$

Because the equation holds, the event that the card is a heart and the event that the card is an ace are independent events.

3.2 Conditional probability

Are students more likely to use marijuana when their parents used drugs? The `drug_use` data set contains a sample of 445 cases with two variables, `student` and `parents`, and is summarized in Table 3.7.[20] The `student` variable is either `uses` or `not`, where a student is labeled as `uses` if she has recently used marijuana. The `parents` variable takes the value `used` if at least one of the parents used drugs, including alcohol.

		parents		
		used	not	Total
student	uses	125	94	219
	not	85	141	226
	Total	210	235	445

Table 3.7: Contingency table summarizing the `drug_use` data set.

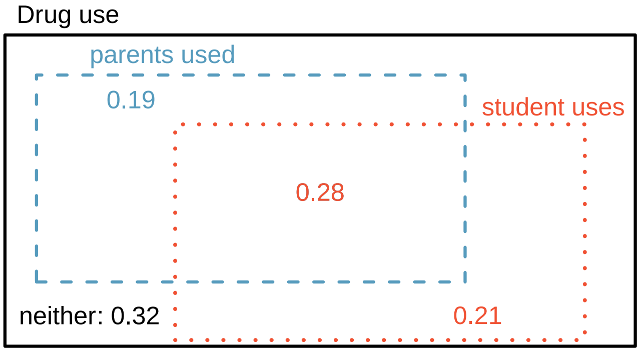


Figure 3.8: A Venn diagram using boxes for the `drug_use` data set.

[20] Ellis GJ and Stone LH. 1979. Marijuana Use in College: An Evaluation of a Modeling Explanation. Youth and Society 10:323-334.

	`parents: used`	`parents: not`	Total
`student: uses`	0.28	0.21	0.49
`student: not`	0.19	0.32	0.51
Total	0.47	0.53	1.00

Table 3.9: Probability table summarizing parental and student drug use.

● **Example 3.31** If at least one parent used drugs, what is the chance their child (`student`) uses?

We will estimate this probability using the data. Of the 210 cases in this data set where **`parents used`**, 125 represent cases where **`student uses`**:

$$P(\texttt{student uses} \text{ given } \texttt{parents used}) = \frac{125}{210} = 0.60$$

● **Example 3.32** A student is randomly selected from the study and she does not use drugs. What is the probability that at least one of her parents used?

If the student does not use drugs, then she is one of the 226 students in the second row. Of these 226 students, 85 had at least one parent who used drugs:

$$P(\texttt{parents used} \text{ given } \texttt{student does not use}) = \frac{85}{226} = 0.376$$

3.2.1 Marginal and joint probabilities

Table 3.9 includes row and column totals for each variable separately in the `drug_use` data set. These totals represent **marginal probabilities** for the sample, which are the probabilities based on a single variable without conditioning on any other variables. For instance, a probability based solely on the `student` variable is a marginal probability:

$$P(\texttt{student uses}) = \frac{219}{445} = 0.492$$

A probability of outcomes for two or more variables or processes is called a **joint probability**:

$$P(\texttt{student uses} \text{ and } \texttt{parents did not use}) = \frac{94}{445} = 0.21$$

It is common to substitute a comma for "and" in a joint probability, although either is acceptable.

Marginal and joint probabilities

If a probability is based on a single variable, it is a *marginal probability.* The probability of outcomes for two or more variables or processes is called a *joint probability.*

We use **table proportions** to summarize joint probabilities for the `drug_use` sample. These proportions are computed by dividing each count in Table 3.7 by 445 to obtain the proportions in Table 3.9. The joint probability distribution of the `parents` and `student` variables is shown in Table 3.10.

Joint outcome	Probability
`parents used, student uses`	0.28
`parents used, student does not use`	0.19
`parents did not use, student uses`	0.21
`parents did not use, student does not use`	0.32
Total	1.00

Table 3.10: A joint probability distribution for the `drug_use` data set.

⊙ **Guided Practice 3.33** Verify Table 3.10 represents a probability distribution: events are disjoint, all probabilities are non-negative, and the probabilities sum to 1.[21]

We can compute marginal probabilities using joint probabilities in simple cases. For example, the probability a random student from the study uses drugs is found by summing the outcomes from Table 3.10 where `student uses`:

$$\begin{aligned} P(\underline{\texttt{student uses}}) &= P(\texttt{parents used, }\underline{\texttt{student uses}}) \\ &\quad + P(\texttt{parents did not use, }\underline{\texttt{student uses}}) \\ &= 0.28 + 0.21 \\ &= 0.49 \end{aligned}$$

3.2.2 Defining conditional probability

There is some connection between drug use of parents and of the student: drug use of one is associated with drug use of the other.[22] In this section, we discuss how to use information about associations between two variables to improve probability estimation.

The probability that a random student from the study uses drugs is 0.49. Could we update this probability if we knew that this student's parents used drugs? Absolutely. To do so, we limit our view to only those 210 cases where parents used drugs and look at the fraction where the student uses drugs:

$$P(\texttt{student uses}\text{ given }\texttt{parents used}) = \frac{125}{210} = 0.60$$

We call this a **conditional probability** because we computed the probability under a condition: `parents used`. There are two parts to a conditional probability, **the outcome of interest** and the **condition**. It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event.

We separate the text inside our probability notation into the outcome of interest and the condition:

$P(A|B)$
Probability of outcome A given B

$$\begin{aligned} &P(\texttt{student uses}\text{ given }\texttt{parents used}) \\ &= P(\texttt{student uses}|\ \texttt{parents used}) = \frac{125}{210} = 0.60 \end{aligned} \qquad (3.34)$$

The vertical bar "|" is read as *given*.

In Equation (3.34), we computed the probability a student uses based on the condition that at least one parent used as a fraction:

[21] Each of the four outcome combination are disjoint, all probabilities are indeed non-negative, and the sum of the probabilities is $0.28 + 0.19 + 0.21 + 0.32 = 1.00$.

[22] This is an observational study and no causal conclusions may be reached.

$$\begin{aligned} &P(\texttt{student uses} \mid \texttt{parents used}) \\ &= \frac{\#\text{ times }\texttt{student uses}\text{ and }\texttt{parents used}}{\#\text{ times }\texttt{parents used}} \\ &= \frac{125}{210} = 0.60 \end{aligned} \tag{3.35}$$

We considered only those cases that met the condition, `parents used`, and then we computed the ratio of those cases that satisfied our outcome of interest, the student uses.

Counts are not always available for data, and instead only marginal and joint probabilities may be provided. For example, disease rates are commonly listed in percentages rather than in a count format. We would like to be able to compute conditional probabilities even when no counts are available, and we use Equation (3.35) as an example demonstrating this technique.

We considered only those cases that satisfied the condition, `parents used`. Of these cases, the conditional probability was the fraction who represented the outcome of interest, `student uses`. Suppose we were provided only the information in Table 3.9 on page 95, i.e. only probability data. Then if we took a sample of 1000 people, we would anticipate about 47% or $0.47 \times 1000 = 470$ would meet our information criterion. Similarly, we would expect about 28% or $0.28 \times 1000 = 280$ to meet both the information criterion and represent our outcome of interest. Thus, the conditional probability could be computed:

$$\begin{aligned} P(\texttt{student uses}| \texttt{ parents used}) &= \frac{\#\ (\texttt{student uses}\text{ and }\texttt{parents used})}{\#\ (\texttt{parents used})} \\ &= \frac{280}{470} = \frac{0.28}{0.47} = 0.60 \end{aligned} \tag{3.36}$$

In Equation (3.36), we examine exactly the fraction of two probabilities, 0.28 and 0.47, which we can write as

$$P(\texttt{student uses}\text{ and }\texttt{parents used}) \quad \text{and} \quad P(\texttt{parents used})$$

The fraction of these probabilities represents our general formula for conditional probability.

Conditional Probability

The conditional probability of the outcome of interest A given condition B is computed as the following:

$$P(A|B) = \frac{P(A\text{ and }B)}{P(B)} \tag{3.37}$$

⊙ **Guided Practice 3.38** (a) Write out the following statement in conditional probability notation: "*The probability a random case has* `parents did not use` *if it is known that* `student does not use`". Notice that the condition is now based on the student, not the parent. (b) Determine the probability from part (a). Table 3.9 on page 95 may be helpful.[23]

[23] (a) $P($`parents did not use`$|$ `student does not use`$)$. (b) Equation (3.37) for conditional probability indicates we should first find $P($`parents did not use` and `student does not use`$) = 0.32$ and $P($`student does not use`$) = 0.51$. Then the ratio represents the conditional probability: $0.32/0.51 = 0.63$.

		inoculated		
		yes	no	Total
result	lived	238	5136	5374
	died	6	844	850
	Total	244	5980	6224

Table 3.11: Contingency table for the `smallpox` data set.

		inoculated		
		yes	no	Total
result	lived	0.0382	0.8252	0.8634
	died	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Table 3.12: Table proportions for the `smallpox` data, computed by dividing each count by the table total, 6224.

⊙ **Guided Practice 3.39** (a) Determine the probability that one of the parents had used drugs if it is known the student does not use drugs. (b) Using the answers from part (a) and Guided Practice 3.38(b), compute

$$P(\texttt{parents used}|\texttt{student does not use})$$
$$+P(\texttt{parents did not use}|\texttt{student does not use})$$

(c) Provide an intuitive argument to explain why the sum in (b) is 1.[24]

⊙ **Guided Practice 3.40** The data indicate that drug use of parents and children are associated. Does this mean the drug use of parents causes the drug use of the students?[25]

3.2.3 Smallpox in Boston, 1721

The `smallpox` data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[26] Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

Each case represents one person with two variables: `inoculated` and `result`. The variable `inoculated` takes two levels: `yes` or `no`, indicating whether the person was inoculated or not. The variable `result` has outcomes `lived` or `died`. These data are summarized in Tables 3.11 and 3.12.

⊙ **Guided Practice 3.41** Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[27]

[24] (a) This probability is $\frac{P(\texttt{parents used and student does not use})}{P(\texttt{student does not use})} = \frac{0.19}{0.51} = 0.37$. (b) The total equals 1. (c) Under the condition the student does not use drugs, the parents must either use drugs or not. The complement still appears to work *when conditioning on the same information.*

[25] No. This was an observational study. Two potential confounding variables include `income` and `region`. Can you think of others?

[26] Fenner F. 1988. *Smallpox and Its Eradication (History of International Public Health, No. 6).* Geneva: World Health Organization. ISBN 92-4-156110-6.

[27] $P(\texttt{result} = \texttt{died} \mid \texttt{not inoculated}) = \frac{P(\texttt{result = died and not inoculated})}{P(\texttt{not inoculated})} = \frac{0.1356}{0.9608} = 0.1411$.

⊙ **Guided Practice 3.42** Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice 3.41?[28]

⊙ **Guided Practice 3.43** The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone `lived` or `died` and also affect whether that person was inoculated?[29]

[28] $P(\texttt{died} \mid \texttt{inoculated}) = \frac{P(\texttt{died and inoculated})}{P(\texttt{inoculated})} = \frac{0.0010}{0.0392} = 0.0255$. The death rate for individuals who were inoculated is only about 1 in 40 while the death rate is about 1 in 7 for those who were not inoculated.

[29] Brief answers: (a) Observational. (b) No, we cannot infer causation from this observational study. (c) Accessibility to the latest and best medical care. There are other valid answers for part (c).

3.2.8 Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

The smallpox data fit this description. We see the population as split by inoculation: yes and no. Following this split, survival rates were observed for each group. This structure is reflected in the tree diagram shown in Figure 3.14. The first branch for inoculation is said to be the **primary** branch while the other branches are **secondary**.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 3.14. This tree diagram splits the smallpox data by inoculation into the yes and no groups with respective marginal probabilities 0.0392 and 0.9608. The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure 3.14 is the probability that lived conditioned on the information that inoculated. We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

$$\begin{aligned} P(\texttt{inoculated} \text{ and } \texttt{lived}) &= P(\texttt{inoculated}) \times P(\texttt{lived} | \ \texttt{inoculated}) \\ &= 0.0392 \times 0.9754 \\ &= 0.0382 \end{aligned}$$

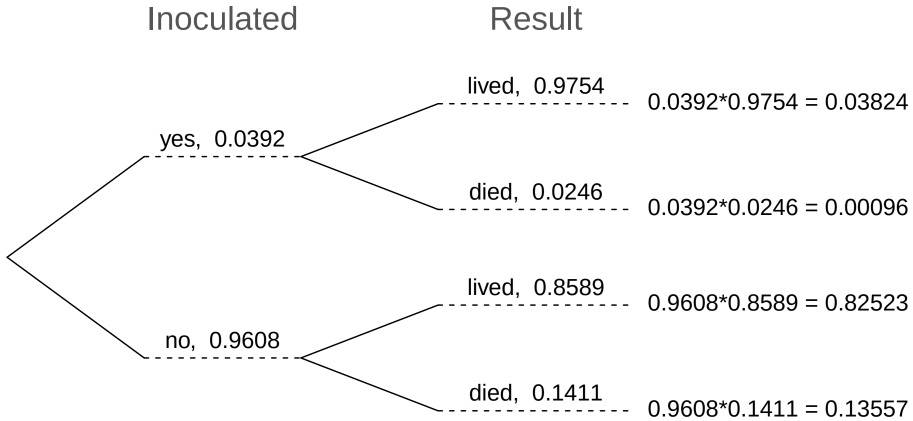


Figure 3.14: A tree diagram of the smallpox data set.

Example 3.60 What is the probability that a randomly selected person who was inoculated died?

This is equivalent to $P(\texttt{died}|\ \texttt{inoculated})$. This conditional probability can be found in the second branch as 0.0246.

Example 3.61 What is the probability that a randomly selected person lived?

There are two ways that a person could have lived: be inoculated *and* live OR not be inoculated *and* live. To find this probability, we sum the two disjoint probabilities:

$$P(\texttt{lived}) = 0.0392 \times 0.9745 + 0.9608 \times 0.8589 = 0.03824 + 0.82523 = 0.86347$$

Guided Practice 3.62 After an introductory statistics course, 78% of students can successfully construct tree diagrams. Of those who can construct tree diagrams, 97% passed, while only 57% of those students who could not construct tree diagrams passed. (a) Organize this information into a tree diagram. (b) What is the probability that a student who was able to construct tree diagrams did not pass? (c) What is the probability that a randomly selected student was able to successfully construct tree diagrams and passed? (d) What is the probability that a randomly selected student passed? [40]

3.2.9 Bayes' Theorem

In many instances, we are given a conditional probability of the form

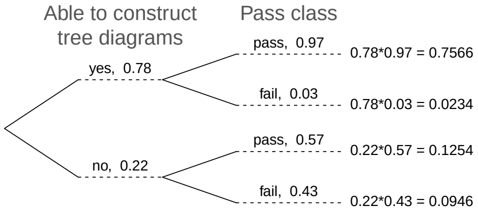
$$P(\text{statement about variable 1}|\ \text{statement about variable 2})$$

but we would really like to know the inverted conditional probability:

$$P(\text{statement about variable 2}|\ \text{statement about variable 1})$$

For example, instead of wanting to know $P(\text{lived} \mid \text{inoculated})$, we might want to know $P(\text{inoculated} \mid \text{lived})$. This is more challenging because it cannot be read directly from the tree diagram. In these instances we use **Bayes' Theorem**. Let's begin by looking at a new example.

[40] (a) The tree diagram is shown to the right. (b) $P(\text{not pass} \mid \text{able to construct tree diagram}) = 0.03$. (c) $P(\text{able to construct tree diagrams and passed}) = P(\text{able to construct tree diagrams}) \times P(\text{passed} \mid \text{able to construct tree diagrams}) = 0.78 \times 0.97 = 0.7566$. (d) $P(\text{passed}) = 0.7566 + 0.1254 = 0.8820$.



● **Example 3.63** In Canada, about 0.35% of women over 40 will be diagnosed with breast cancer in any given year. A common screening test for cancer is the mammogram, but this test is not perfect. In about 11% of patients with breast cancer, the test gives a **false negative**: it indicates a woman does not have breast cancer when she does have breast cancer. Similarly, the test gives a **false positive** in 7% of patients who do not have breast cancer: it indicates these patients have breast cancer when they actually do not.[41] If we tested a random woman over 40 for breast cancer using a mammogram and the test came back positive – that is, the test suggested the patient has cancer – what is the probability that the patient actually has breast cancer?

We are given sufficient information to quickly compute the probability of testing positive if a woman has breast cancer ($1.00 - 0.11 = 0.89$). However, we seek the inverted probability of cancer given a positive test result:

$$P(\text{has BC} \mid \text{mammogram}^+)$$

Here, "has BC" is an abbreviation for the patient actually having breast cancer, and "mammogram$^+$" means the mammogram screening was positive, which in this case means the test suggests the patient has breast cancer. (Watch out for the non-intuitive medical language: a *positive* test result suggests the possible presence of cancer in a mammogram screening.) We can use the conditional probability formula from the previous section: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. Our conditional probability can be found as follows:

$$P(\text{has BC} \mid \text{mammogram}^+) = \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)}$$

The probability that a mammogram is positive is as follows.

$$P(\text{mammogram}^+) = P(\text{has BC and mammogram}^+) + P(\text{no BC and mammogram}^+)$$

A tree diagram is useful for identifying each probability and is shown in Figure 3.15.

[41] The probabilities reported here were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

Using the tree diagram, we find that

$$\begin{aligned}
&P(\text{has BC} \mid \text{mammogram}^{+})\\
&\quad= \frac{P(\text{has BC and mammogram}^{+})}{P(\text{has BC and mammogram}^{+}) + P(\text{no BC and mammogram}^{+})}\\
&\quad= \frac{0.0035(0.89)}{0.0035(0.89) + 0.9965(0.07)}\\
&\quad= \frac{0.00312}{0.07288} \approx 0.0428
\end{aligned}$$

That is, even if a patient has a positive mammogram screening, there is still only a 4% chance that she has breast cancer.

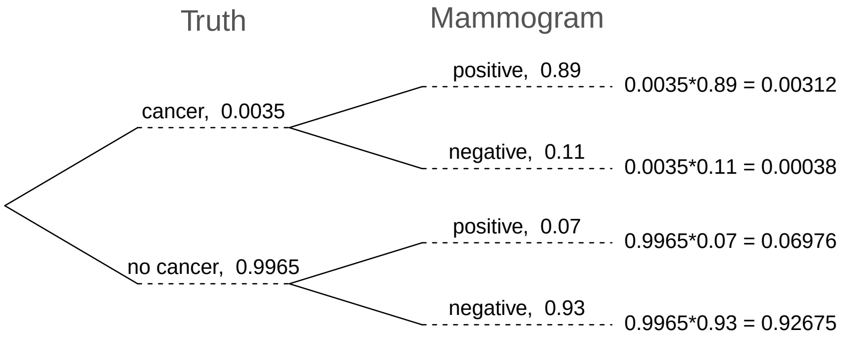


Figure 3.15: Tree diagram for Example 3.63, computing the probability a random patient who tests positive on a mammogram actually has breast cancer.

Example 3.63 highlights why doctors often run more tests regardless of a first positive test result. When a medical condition is rare, a single positive test isn't generally definitive.

Consider again the last equation of Example 3.63. Using the tree diagram, we can see that the numerator (the top of the fraction) is equal to the following product:

$$P(\text{has BC and mammogram}^{+}) = P(\text{mammogram}^{+} \mid \text{has BC})P(\text{has BC})$$

The denominator – the probability the screening was positive – is equal to the sum of probabilities for each positive screening scenario:

$$P(\underline{\text{mammogram}^{+}}) = P(\underline{\text{mammogram}^{+}} \text{ and no BC}) + P(\underline{\text{mammogram}^{+}} \text{ and has BC})$$

In the example, each of the probabilities on the right side was broken down into a product of a conditional probability and marginal probability using the tree diagram.

$$\begin{aligned}
P(\text{mammogram}^{+}) &= P(\text{mammogram}^{+} \text{ and no BC}) + P(\text{mammogram}^{+} \text{ and has BC})\\
&= P(\text{mammogram}^{+} \mid \text{no BC})P(\text{no BC})\\
&\qquad + P(\text{mammogram}^{+} \mid \text{has BC})P(\text{has BC})
\end{aligned}$$

We can see an application of Bayes' Theorem by substituting the resulting probability expressions into the numerator and denominator of the original conditional probability.

$$P(\text{has BC}|\text{ mammogram}^+)$$
$$= \frac{P(\text{mammogram}^+|\text{ has BC})P(\text{has BC})}{P(\text{mammogram}^+|\text{ no BC})P(\text{no BC}) + P(\text{mammogram}^+|\text{ has BC})P(\text{has BC})}$$

Bayes' Theorem: inverting probabilities

Consider the following conditional probability for variable 1 and variable 2:

$$P(\text{outcome } A_1 \text{ of variable 1}|\text{ outcome } B \text{ of variable 2})$$

Bayes' Theorem states that this conditional probability can be identified as the following fraction:

$$\frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)} \tag{3.64}$$

where A_2, A_3, ..., and A_k represent all other possible outcomes of the first variable.

Bayes' Theorem is just a generalization of what we have done using tree diagrams. The formula can be memorized. If not, it is important to be able to derive the formula quickly with a tree diagram:

- The numerator identifies the probability of getting both A_1 and B.
- The denominator is the overall probability of getting B.

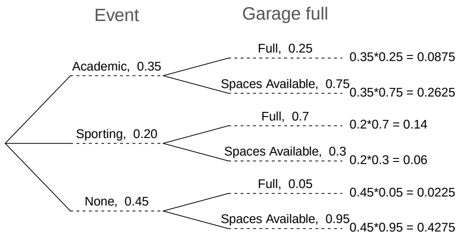
The bottom component (the denominator) of the fraction often appears long and complicated. However, it is equivalent to what we did numerically using tree diagrams: traverse each branch of the tree diagram that ends with event B.

⊙ **Guided Practice 3.65** Jose visits campus every Thursday evening. However, some days the parking garage is full, often due to college events. There are academic events on 35% of evenings, sporting events on 20% of evenings, and no events on 45% of evenings. When there is an academic event, the garage fills up about 25% of the time, and it fills up 70% of evenings with sporting events. On evenings when there are no events, it only fills up about 5% of the time. If Jose comes to campus and finds the garage full, what is the probability that there is a sporting event? Use a tree diagram to solve this problem.[42]

[42] The tree diagram, with three primary branches, is shown to the right. We want

$$P(\text{sporting event}|\text{garage full})$$
$$= \frac{P(\text{sporting event and garage full})}{P(\text{garage full})}$$
$$= \frac{0.14}{0.0875 + 0.14 + 0.0225} = 0.56.$$

If the garage is full, there is a 56% probability that there is a sporting event.



Appendix A

End of chapter exercise solutions

2 Summarizing data

2.3 (a) 1/linear and 3/nonlinear. (b) 4/some curvature (nonlinearity) may be present on the right side. “Linear” would also be acceptable for the type of relationship for plot 4. (c) 2.

2.5 (a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean. Use a weight of 24 for the old mean and 1 for the new mean: $(24 \times 74 + 1 \times 64)/(24 + 1) = 73.6$. There are other ways to solve this exercise that do not use a weighted mean. (c) The new score is more than 1 standard deviation away from the previous mean, so increase.

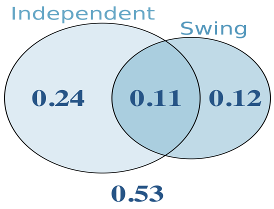
3 Probability

3.1 (a) False. These are independent trials. (b) False. There are red face cards. (c) True. A card cannot be both a face card and an ace.

3.3 (a) 10 tosses. Fewer tosses mean more vari- ability in the sample fraction of heads, mean- ing there’s a better chance of getting at least 60% heads. (b) 100 tosses. More flips means the observed proportion of heads would often be closer to the average, 0.50, and therefore also above 0.40. (c) 100 tosses. With more flips, the observed proportion of heads would often be closer to the average, 0.50. (d) 10 tosses. Fewer flips would increase variability in the fraction of tosses that are heads.

3.5 (a) $0.5^{10} = 0.00098$. (b) $0.5^{10} = 0.00098$. (c) $P(\text{at least one tails}) = 1 - P(\text{no tails}) = 1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

3.7 (a) No, there are voters who are both politically Independent and also swing voters. (b) Venn diagram below:



(c) 24%. (d) Add up the corresponding dis- joint sections in the Venn diagram: $0.24 + 0.11 + 0.12 = 0.47$. Alternatively, use the Gen- eral Addition Rule: $0.35 + 0.23 - 0.11 = 0.47$. (e) $1 - 0.47 = 0.53$. (f) $P(\text{Independent}) \times P(\text{swing}) = 0.35 \times 0.23 = 0.08$, which does not equal P(Independent and swing) = 0.11, so the events are dependent. If you stated that this difference might be due to sampling variability in the survey, that answer would also be rea- sonable (we’ll dive into this topic more in later chapters).

3.9 (a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint (unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c)). (b) They are probably not independent: If you study together, your study habits would be related, which suggests your course performances are also related. (c) No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.

3.11 (a) 0.16 + 0.09 = 0.25. (b) 0.17 + 0.09 = 0.26. (c) Assuming that the education level of the husband and wife are independent: 0.25 X 0.26 = 0.065. You might also notice we actually made a second assumption: that the decision to get married is unrelated to education level. (d) The husband/wife independence assumption is probably not reasonable, because people often marry another person with a comparable level of education. We will leave it to you to think about whether the second assumption noted in part (c) is reasonable.

3.13 (a) Invalid. Sum is greater than 1. (b) Valid. Probabilities are between 0 and 1, and they sum to 1. In this class, every student gets a C. (c) Invalid. Sum is less than 1. (d) In- valid. There is a negative probability. (e) Valid. Probabilities are between 0 and 1, and they sum to 1. (f) Invalid. There is a negative probability.

3.15 (a) No, but we could if A and B are inde- pendent. (b-i) 0.21. (b-ii) $0.3+0.7-0.21 = 0.79$. (b-iii) Same as $P(A)$: 0.3. (c) No, because $0.1 \neq 0.21$, where 0.21 was the value computed un- der independence from part (a). (d) $P(\text{A}|\text{B}) = 0.1/0.7 = 0.143$.

3.17 (a) $0.60 + 0.20 - 0.18 = 0.62$. (b) $0.18/0.20 = 0.90$. (c) $0.11/0.33 \approx 0.33$. (d) No, otherwise the final answers of parts (b) and (c) would have been equal. (e) $0.06/0.34 \approx 0.18$.

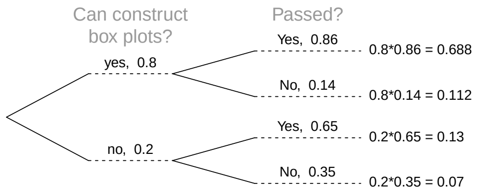
3.19 (a) $162/248 = 0.65$. (b) $181/252 = 0.72$ (c) Under the assumption of a dating choices be- ing independent of hamburger preference, which on the surface seems reasonable: $0.65 \times 0.72 = 0.468$. (d) $(252 + 6 - 1)/500 = 0.514$

3.21 (a) 0.3. (b) 0.3. (c) 0.3. (d) $0.3 \times 0.3 = 0.09$. (e) Yes, the population that is being sam- pled from is identical in each draw.

3.23 (a) 2/9. (b) $3/9 = 1/3$. (c) $(3/10) \times (2/9) \approx 0.067$. (d) No. In this small population of marbles, removing one marble meaningfully changes the probability of what might be drawn next.

3.25 For 1 leggings (L) and 2 jeans (J), there are three possible orderings: LJJ, JLJ, and JJL. The probability for LJJ is $(5/24) \times (7/23) \times (6/22) = 0.0173$. The other two orderings have the same probability, and these three possible orderings are disjoint events. FInal answer: 0.0519.

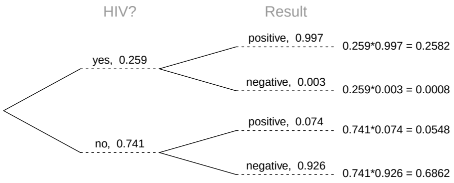
3.27 (a) The tree diagram:



(b) $P(\text{can construct}|\text{pass}) = \frac{P(\text{can construct and pass})}{P(\text{pass})}$

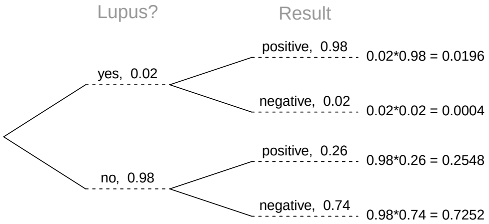
$\frac{0.8\times 0.86}{0.8\times 0.86\ +\ 0.2\times 0.65} = \frac{0.688}{0.818} \approx 0.84.$

3.29 First draw a tree diagram:



Then compute the probability: $P(HIV|+) = \frac{P(HIV \ and \ +)}{P(+)} = \frac{0.259 \times 0.997}{0.259 \times 0.997 + 0.741 \times 0.074} = \frac{0.2582}{0.3131} = 0.8247$.

3.31 A tree diagram of the situation:



$P(\text{lupus}|\text{positive}) = \frac{P(\text{lupus and positive})}{P(\text{positive})} = \frac{0.0196}{0.0196+0.2548} = 0.0714$. Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. While House is not exactly right – it is possible that the pa- tient has lupus – his implied skepticism is war- ranted.

3.33 (a) $\overset{Anna}{1/5} \times \overset{Ben}{1/4} \times \overset{Carl}{1/3} \times \overset{Damian}{1/2} \times \overset{Eddy}{1/1} =$ $1/5! = 1/120$. (b) Since the probabilities must add to 1, there must be 5! - 120 possible orderings. (c) 8! = 40,320.

3.35 (a) Yes. The conditions are satisfied: independence, fixed number of trials, either success or failure for each trial, and probability of success being constant across trials. (b) 0.0200. (c) 0.200. (d) 0.0024 + 0.0284 +0.1323 = 0.1631. (e) 1 - 0.0024 = 0.9976.

3.37 (a) $\mu = 35$, $\sigma = 3.24$. (b) Yes. $Z = 3.09$. Since 45 is more than 2 standard deviations from the mean, it would be considered unusual. Note that the normal model is not required to apply this rule of thumb. (c) Using a normal model: 0.0010. This does indeed appear to be an un- usual observation. If using a normal model with a 0.5 correction, the probability would be calcu- lated as 0.0017.

3.39 (a) The table below summarizes the probability model:

Event	X	P(X)	X · P(X)	$(X - E(X))^2$	$(X - E(X))^2 \cdot P(X)$
3 hearts	50	$\frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} = 0.0129$	0.65	$(50 - 3.59)^2 = 2154.1$	$2154.1 \times 0.0129 = 27.9$
3 blacks	25	$\frac{26}{52} \times \frac{25}{51} \times \frac{24}{50} = 0.1176$	2.94	$(25 - 3.59)^2 = 458.5$	$458.5 \times 0.1176 = 53.9$
Else	0	$1 - (0.0129 + 0.1176) = 0.8695$	0	$(0 - 3.59)^2 = 12.9$	$12.9 \times 0.8695 = 11.2$
			E(X) = \$3.59		$V(X) = 93.0$ $SD(X) = \sqrt{V(X)} = 9.64$

(b) $E(X-5) = E(X)-5 = 3.59-5 = -\$1.41$. The standard deviation is the same as the standard deviation of X: \$9.64. (c) No. The expected earnings is negative, so on average you would lose money playing the game.

3.41

Event	X	$P(X)$	$X \cdot P(X)$
Boom	0.18	$\frac{1}{3}$	$0.18 \times \frac{1}{3} = 0.06$
Normal	0.09	$\frac{1}{3}$	$0.09 \times \frac{1}{3} = 0.03$
Recession	-0.12	$\frac{1}{3}$	$-0.12 \times \frac{1}{3} = -0.04$
			$E(X) = 0.05$

The expected return is a 5% increase in value for a single year.

3.43 (a) Expected: -\$0.16. Variance: 8.95. SD: \$2.99. (b) Expected: -\$0.16. SD: \$1.73. (c) Expected values are the same, but the SDs differ. The SD from the game with tripled win- nings/losses is larger, since the three indepen- dent games might go in different directions (e.g. could win one game and lose two games). So the three independent games is lower risk, but in this context, it just means we are likely to lose a more stable amount since th eexpected value is still negative.

3.45 A fair game has an expected value of zero: $\$5 \times 0.46 + x \times 0.54 = 0$. Solving for x: -\$4.26. You would bet \$4.26 for the Padres to make the game fair.

3.47 (a) Expected: \$3.90. SD: \$0.34. (b) Ex- pected: \$27.30. SD: \$0.89. If you computed part (b) using part (a), you should have ob- tained an SD of \$0.90.

3.49 Approximate answers are OK. Answers are only estimates based on the sample. (a) $(29 + 32)/144 = 0.42$. (b) $21/144 = 0.15$. (c) $(26 + 12 + 15)/144 = 0.37$.